

Exercise 4

Lecture 7: Schema Integration

Scalable Data Engineering

1.1 Are the following statements true or false? Give reasons for your answer.

- a) Schema matching is automated schema mapping.
- b) The characteristics of an integrated schema, i.e., completeness, correctness, minimality, and understandability, can be proven algorithmically.
- c) Several distance measures (matchers) should be used to automatically detect similarities.

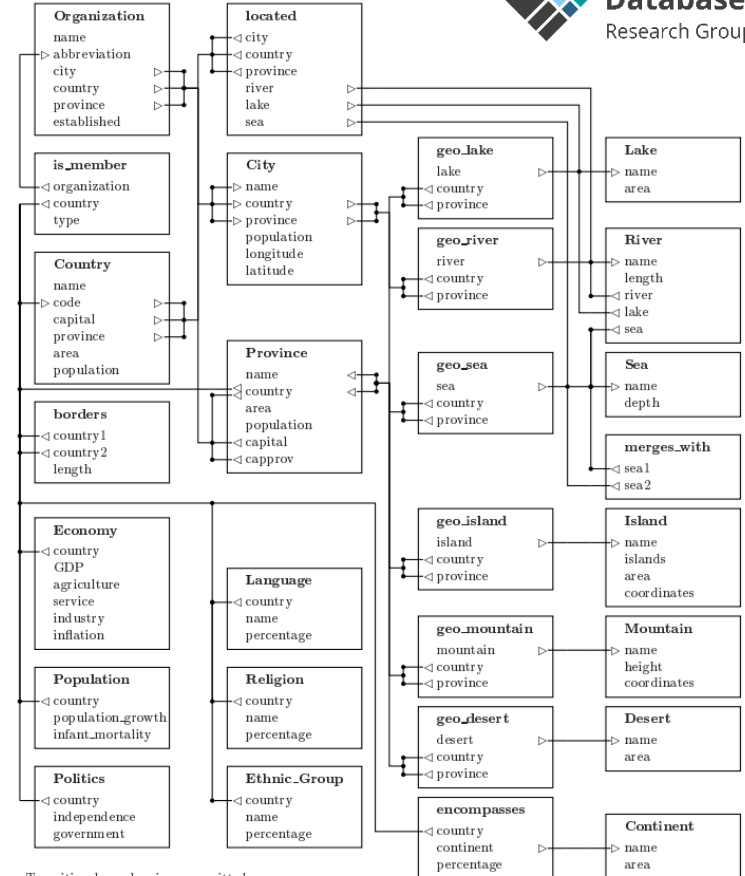
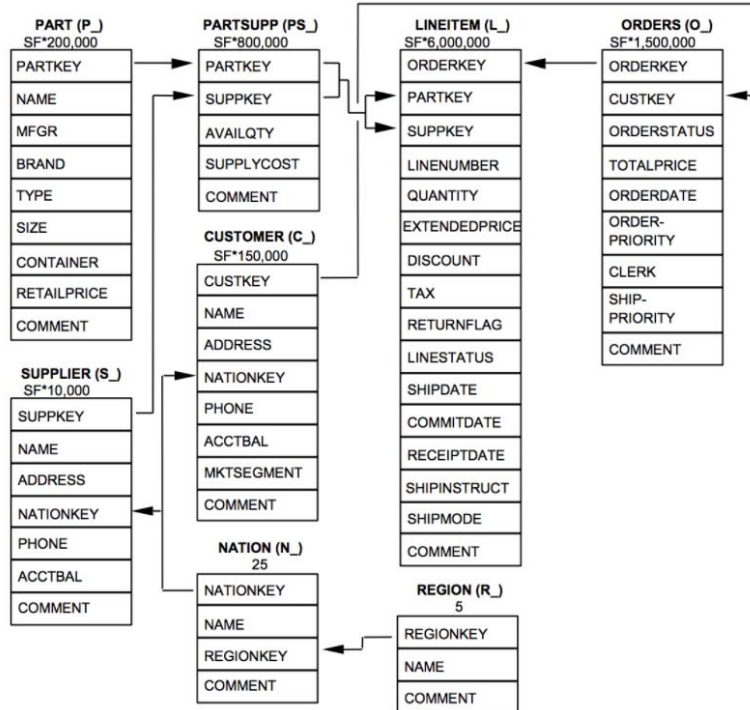
1.2 Match the following schemas. Use the stable marriage algorithm with bigram similarity for scoring.

- $sim(w_1, w_2) = \frac{2 \cdot |bigrams(w_1) \cap bigrams(w_2)|}{|bigrams(w_1)| + |bigrams(w_2)|}$ e.g. $sim(hat, hut) = \frac{2 \cdot |\{h, t\}|}{|\{h, ha, at, t\}| + |\{h, hu, ut, t\}|} = \frac{2 \cdot 2}{4 + 4} = 0.5$



Schema Matching

TPCH and Mondial Schema



Transitive dependencies are omitted.

1.3 Enrich the TPCB Schema using geo information

- a) Import the mondial database by opening and executing mondial.ddl first and mondial.import second using pg_admin. Pay attention to the newly created schema when accessing mondial data.
 - Import scripts here: <https://cloudstore.zih.tu-dresden.de/index.php/s/4xfXiNndcJgtKrN>
- b) Compare the mondial and the TPCB schema, what could be a possible intersection regarding the modeled information?
- c) Since tables in the mondial database do not have unique ID columns, start to add a unique ID to the table/relation on the mondial side.
- d) Now, on the supplier table/relation of the TPCB schema, connect to the newly created ID column.